T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	合成音声のスタイル制御における話者適応の評価
Title(English)	A speaker adaptation technique for MRHSMM-based style control of synthetic speech
著者(和文)	井澤信介, 能勢 隆, 山岸順一, 小林隆夫
Authors(English)	Shinsuke Izawa, Takashi Nose, Junichi Yamagishi, Takao Kobayashi
出典(和文)	日本音響学会2007年春季研究発表会講演論文集, Vol. , No. , pp. 269-270
Citation(English)	Proceedings of the ASJ 2007 Spring Meeting, Vol. , No. , pp. 269-270
発行日 / Pub. date	2007, 3

Powered by T2R2 (Science Tokyo Research Repository)

合成音声のスタイル制御における話者適応の評価*

◎井澤信介, 能勢隆, 山岸順一, 小林隆夫 (東工大)

1 はじめに

音声合成においてより人間らしい機能を実現するために，様々な発話様式・感情表現 (スタイル) を含んだ音声を任意の話者の声で容易に実現することが望まれている．その手法の一つとして，我々は隠れセミマルコフモデル (HSMM) に基づく音声合成システム [1] を拡張した重回帰 HSMM に基づくスタイル制御法を提案した [2]．重回帰 HSMM に基づくスタイル制御では複数のスタイルを同時に単一のモデルで表現し，スタイルの表出・強調度合を低次元のベクトルで表現することにより，直観的に合成音声のスタイルの表出度合を制御することが可能である．さらに，学習データ収録のコスト低減のために，HMM に基づく音声認識などで広く用いられている最尤線形回帰による話者適応手法を重回帰 HSMM に導入することで，少量の目標話者のデータから目標話者の合成音声のスタイル制御を可能にする話者適応手法を提案した [3]．本論文では，この手法の主観評価実験を行い，提案手法の有効性を検討する．

2 重回帰 HSMM を用いたスタイル制御における話者適応

重回帰 HSMM [2] では状態 i の出力分布 $b_i(\boldsymbol{o})$ 及び状態継続長分布 $p_i(d)$ の平均が次のような重回帰式で表されると仮定する．

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} \tag{1}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{2}$$

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{3}$$

ここで $\boldsymbol{v}$ はスタイルの表出・強調度合を表した低次元のベクトル (スタイルベクトル) であり，$\boldsymbol{H}_{b_i}$ および $\boldsymbol{H}_{p_i}$ はそれぞれ $M \times (L+1)$ 次元，$1 \times (L+1)$ 次元の回帰行列，M は特徴ベクトルの次元である．なお，$b_i(\boldsymbol{o})$ および $p_i(d)$ は単一ガウス分布を仮定している．モデルパラメータの学習は EM アルゴリズムによる最尤推定で行い [2] ，音声合成時は，学習で得られた回帰行列に対し所望のスタイルの表出・強調度合に対応するスタイルベクトルを与えることにより，出力分布および状態継続長分布の平均を計算し，音声パラメータを生成する．

今，あるスタイル s において変換元となる話者のモデルと目標話者のモデルの間に以下のアフィン変換の関係が成り立つと仮定する．

$$\hat{\boldsymbol{\mu}}_i^{(s)} = \boldsymbol{A}_{b_i}^{(s)}\boldsymbol{\mu}_i^{(s)} + \boldsymbol{b}_{b_i}^{(s)} \tag{4}$$

ここで $\hat{\boldsymbol{\mu}}_i^{(s)}$ および $\boldsymbol{\mu}_i^{(s)}$ はそれぞれ目標話者，元話者の平均パラメータを表す．重回帰 HSMM では，この平均パラメータは回帰行列 $\boldsymbol{H}_{b_i}$ と $\boldsymbol{\xi} = [1, \boldsymbol{v}^\top]^\top$ により表現され，

$$\hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi}^{(s)} = \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}\boldsymbol{\xi}^{(s)} + \boldsymbol{b}_{b_i}^{(s)} \tag{5}$$

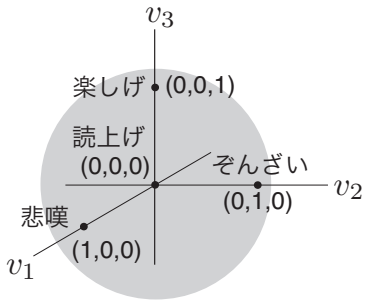

Fig. 1　スタイル空間

となる．ここで，回帰行列 $\boldsymbol{A}_{b_i}$ は各スタイルに共通とし，バイアス項 $\boldsymbol{b}_{b_i}^{(s)}$ がスタイルベクトルを用いて

$$\boldsymbol{b}_{b_i}^{(s)} = \boldsymbol{B}_{b_i}\boldsymbol{\xi}^{(s)} \tag{6}$$

と表されるとすると，式 (5) は

$$\hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi}^{(s)} = \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}\boldsymbol{\xi}^{(s)} + \boldsymbol{B}_{b_i}\boldsymbol{\xi}^{(s)} \tag{7}$$

$$= (\boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i} + \boldsymbol{B}_{b_i})\,\boldsymbol{\xi}^{(s)} \tag{8}$$

となり，結果として重回帰 HSMM における話者適応は，元話者の回帰行列 $\boldsymbol{H}_{b_i}$ を次のように線形変換することにより実現される．

$$\hat{\boldsymbol{H}}_{b_i} = \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i} + \boldsymbol{B}_{b_i} \tag{9}$$

なお，状態継続長分布の回帰行列 $\boldsymbol{H}_{p_i}$ の場合も同様にして求められ，この変換行列は EM アルゴリズムを用いた最尤推定法により求められる [3].

3 実験

3.1 実験条件

重回帰 HSMM の学習には，男性話者 MMI と女性話者 FTY が ATR 音韻バランス文をそれぞれ「読上げ」,「悲嘆」,「楽しげ」,「ぞんざい」のスタイルで発声したデータ [4] を用いた．

サンプリングレート 16kHz の音声信号を，フレーム長 25ms，フレーム周期 5ms のブラックマン窓を用いてメルケプストラム分析し，0 次から 24 次のメルケプストラムを求めた．得られたメルケプストラムと対数基本周波数，及びこれらの Δ および Δ^2 パラメータからなる 78 次元のベクトルを特徴ベクトルとし，5 状態の left-to-right 重回帰 HSMM によりモデル化した．各話者の重回帰 HSMM(話者依存重回帰 HSMM) の学習には，各スタイル 450 文章，計 1800 文章を用いた．本研究では，元話者を FTY，目標話者を MMI とし，FTY の話者依存重回帰 HSMM モデルに対し，学習文章に含まれている各スタイル 50 文章，計 200 文章の MMI のデータを用いて適応を行った．

*A speaker adaptation technique for MRHSMM-based style control of synthetic speech, by IZAWA, Shinsuke, NOSE, Takashi, YAMAGISHI, Junichi, and KOBAYASHI, Takao (Tokyo Institute of Technology).

Table 1　スタイル再現性の評価

(a) Adapted MRHSMM

スタイルおよび	Classification Rate (%)				
スタイルベクトル	読上げ	悲嘆	ぞんざい	楽しげ	その他
読上げ (0,0,0)	90.0	0.0	4.3	1.4	4.3
悲嘆 (1,0,0)	15.7	80.0	2.9	0.0	1.4
ぞんざい (0,1,0)	7.1	0.0	90.0	0.0	2.9
楽しげ (0,0,1)	11.4	0.0	1.4	87.1	0.0

(b) SD MRHSMM

スタイルおよび	Classification Rate (%)				
スタイルベクトル	読上げ	悲嘆	ぞんざい	楽しげ	その他
読上げ (0,0,0)	100.0	0.0	0.0	0.0	0.0
悲嘆 (1,0,0)	1.4	97.1	0.0	0.0	1.4
ぞんざい (0,1,0)	4.3	1.4	94.3	0.0	0.0
楽しげ (0,0,1)	4.3	0.0	0.0	95.7	0.0

スタイル空間は図 1 に示すような三次元空間である．学習データと適応データのスタイルベクトルは，「読上げ」(0,0,0),「悲嘆」(1,0,0),「楽しげ」(0,1,0),「ぞんざい」(0,0,1) としている．重回帰 HSMM における話者適応の変換行列は，静的特徴量と Δ および Δ^2 パラメータからなるブロック対角行列である．また提案手法との比較のために，目標話者の各スタイル 450 文章で学習したスタイル依存 HSMM [4] を用いる．

3.2　スタイルの再現性の評価

主観評価実験の被験者は成人男性 7 名であり，学習データと適応データに含まれない 53 文章からランダムに選択された 10 文章を用いた．

まず，適応を行った重回帰 HSMM(Adapted MRHSMM) に対し，学習時に使用したスタイルベクトルを与えて生成した合成音声のスタイルの識別評価を行った．比較のために，目標話者の話者依存重回帰 HSMM(SD MRHSMM) モデルからの合成音声についても同様の実験を行った．被験者には評価音声をランダムに提示し，学習時のスタイルに「その他」を加えた 5 つの中から選択させた．表 1 に実験結果を示す．

この結果から，適応による重回帰 HSMM は話者依存重回帰 HSMM には及ばないものの，概ね意図したスタイルに判定されたことが分かる．

3.3　話者性の再現性の評価

次に，話者性の再現性の実験を行った．被験者に，元話者および目標話者のスタイル依存 HSMM による合成音声，3.2 節と同様の評価音声を順に聞かせ，評価音声を「5：目標話者とほぼ同じ」「4：目標話者に近い」「3：どちらともいえない」「2：元話者に近い」「1：元話者とほぼ同じ」の 5 段階で評価させた．また，比較のために HSMM に基づく MLLR による話者適応 [5] を用いて元話者の各スタイル依存モデルから目標話者のスタイルへスタイル毎に話者適応した合成音声も同様に評価した．図 2 に 95% 信頼区間を付記した結果を示す．

この結果から，重回帰 HSMM に基づく適応では，全てのスタイルにおいてスタイル依存 HSMM に基づく適応とほぼ同等の話者性の再現性が得られることが確認された．

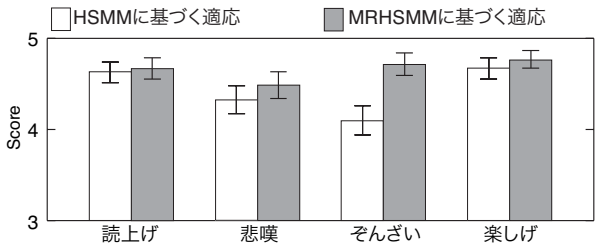

Fig. 2　スタイルごとの話者性の再現性の評価

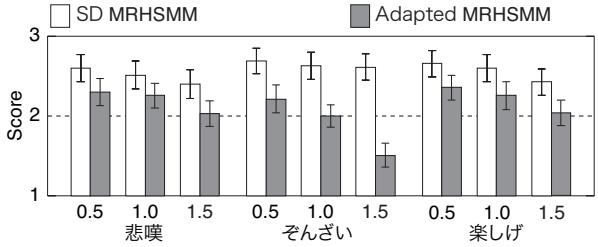

Fig. 3　スタイル制御における自然性の評価

3.4　スタイル制御における自然性の評価

最後に，スタイルベクトルを変化させた場合の合成音声の自然性の評価実験を行った．評価は,「読上げ」を除いた各スタイルについて，合成時のスタイルベクトルをそれぞれ単独で 0.5〜1.5 の間を 0.5 間隔で 3 段階に変化させ,「3：良い」「2：許容できる」「1：悪い」の 3 段階で評価した．比較のために，話者依存重回帰 HSMM に対して同様にスタイルベクトルを変化させて生成した合成音声についても評価を行った．図 3 に，95% 信頼区間を付記した実験結果を示す．

この結果から，特にスタイルベクトルが 1.0 よりも大きいとき，適応は合成音声の自然性が低下していることが分かる．

4　むすび

本論文では，重回帰 HSMM に基づくスタイル制御のための話者適応手法の評価を行った．主観評価実験により，重回帰 HSMM に基づく適応は概ね意図したスタイルを再現でき，HSMM に基づく適応とほぼ同等の話者性の再現性が得られることを確認した．しかし各スタイルを強調した場合の合成音声については自然性が低下することから，このような場合における品質の改善が今後の課題である．

参考文献

[1] Zen, *et al.*, “Hidden semi-Markov model based speech synthesis,” Proc. ICSLP 2004, 1397–1400, 2004.

[2] 能勢他 , “重回帰 HSMM を用いた合成音声のスタイル制御,” 信学技報, SP2005-160, 61–66, 2006.

[3] 井澤他 , “合成音声のスタイル制御における話者適応の検討,” 音響論 (秋), 2006.

[4] Yamagishi, *et al.*, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. & Syst. E88-D(3), 502–509, 2005.

[5] Yamagishi, *et al.*, “MLLR adaptation for hidden semi-Markov model based speech synthesis,” Proc. ICSLP 2004, 1213–1216, 2004.